УДК 004.4'422, 004.432.2, 004.4'418, 51-37

KERNELGEN — ПРОТОТИП РАСПАРАЛЛЕЛИВАЮЩЕГО КОМПИЛЯТОРА C/FORTRAN ДЛЯ GPU NVIDIA НА ОСНОВЕ ТЕХНОЛОГИЙ LLVM[1]

Н.Н. Лихогруд, Д.Н. Микушин

Проект KernelGen (`http://kernelgen.org/`) имеет цель создать на основе современных открытых технологий компилятор Fortran и C для автоматического портирования приложений на GPU без модификации их исходного кода. Анализ параллелизма в KernelGen основан на инфраструктуре LLVM/Polly и CLooG, модифицированной для генерации GPU-ядер и alias-анализе времени исполнения. PTX-ассемблер для GPU NVIDIA генерируется с помощью бекенда NVPTX. Благодаря интеграции LLVM-части с GCC с помощью плагина DragonEgg и модифицированного компоновщика, KernelGen способен, при полной совместимости с компилятором GCC, генерировать исполняемые модули, содержащие одновременно CPU- и GPU-варианты машинного кода. В сравнительных тестах с OpenACC-компилятором PGI KernelGen демонстрирует бóльшую гибкость по ряду возможностей, обеспечивая при этом сравнимый или до 60 % более высокий уровень производительности.

Ключевые слова: GPU, LLVM, OpenACC, JIT-компиляция, выпуклый анализ.

Введение

Широкое использование GPU в кластерных вычислительных системах требует массовой адаптации множества сложных приложений. Программные модели CUDA и OpenCL достаточно хорошо подходят для небольших программ с ярко выраженным вычислительным ядром. Однако для сложных приложений, состоящих из множества отдельных блоков, таких как математические модели, сложность настройки взаимодействия оригинального кода и кода для GPU многократно возрастает. Многие компании и научные группы по-прежнему откладывают портирование своих приложений, так как это приводит к возрастанию издержек на сопровождение и развитие нескольких различных версий одной и той же функциональности для CPU и GPU. Увеличить эффективность портирования призваны следующие типы технологий разработки:

- **Директивные расширения существующих языков высокого уровня с ручным управлением параллелизмом, по аналогии с OpenMP.** Данный тип технологий расширяет языки специальными директивами, с помощью которых пользователь может пометить части кода, предназначенные для выполнения на GPU. Основываясь на данной информации, компилятор автоматически генерирует гибридный исполняемый файл. Для стандартизации набора директив в языках C/C++/Fortran коммерческими разработчиками подобных решений созданы консорциумы OpenACC [7] и OpenHMPP [8]. Аналогичный набор директив, но уже для ускорителей архитектуры Many Integrated Core (MIC) развивается компанией Intel [9]. В рамках систем F2C-ACC [10] и САПФОР [11] предложены наборы директив для преобразования исходного кода на языке Fortran в гибридную форму, причем САПФОР проводит распараллеливание как на уровне GPU, так и на уровне нескольких GPU-узлов MPI-кластера.

[1]Статья рекомендована к публикации программным комитетом Международной научной конференции «Параллельные вычислительные технологии - 2013».

В целом, несмотря на бóльшую гибкость, директивные расширения все же требуют значительного участия программиста в организации корректных и эффективных вычислений. Компиляторы некоторых из приведенных директивных расширений реализуют проверку параллельности циклов и непротиворечивости директив, в других такая проверка отсутствует. Часто возникают ситуации, в которых компилятор слишком «осторожен» при принятии решений на основе внутреннего анализе циклов и производит распараллеливание только при наличии дополнительных указаний от пользователя. Большинство директивных расширений не поддерживают генерацию GPU-ядер для циклов, в которых присутствуют вызовы функций из других модулей компиляции или библиотек, что существенно ограничивает применимость подобных технологий в больших проектах.

- **Специализированные языки (domain-specific languages, DSL), со встроенными средствами параллелизма, ориентированные на определенный класс задач.** В последние годы было предложено множество различных DSL- и Embedded DSL-языков со встроенными средствами параллелизма. Их основная идея состоит в том, чтобы приблизить средства языка к характерным объектам и действиям задачи, в то же время исключив из языка конструкции, привязывающие реализацию к конкретной архитектуре. Обработка возникающего нового уровня абстракции производится компилятором или source-to-source процессором, генерирующим код для всех целевых архитектур. Так, в работе [1] предложен Си-подобный DSL-язык для выражения вычислений на сетках с учетом начальных и граничных условий, а также соответствующий компилятор с бекендами для различных CPU (SSE, AVX). В работе [4] аналогичная задача решается при помощи eDSL, основанного на шаблонах C++. Поддерживается генерация кода для CPU и GPU NVIDIA. Другой eDSL на основе C++ – Halide [5], с поддержкой x86-64/SSE, ARM v7/NEON и GPU NVIDIA, нацелен на эффективную реализацию методов обработки изображений. К классу DSL/eDSL можно отнести систему Nemerle Unified Device Architecture (NUDA) [6], позволяющую создавать новые расширения языка Nemerle и соответствующие плагины для компилятора. Тестирование DSL/eDSL как правило проводится в сравнении с программами, написанными вручную, что не позволяет судить о том, насколько существенен может быть выигрыш в эффективности DSL-языков по сравнению с директивными расширениями. Кроме того специализация ограничивает конкурентную среду, так как у каждого языка как правило существует только один разработчик. Глубокое сравнительное тестирование затруднено необходимостью реализации бенчмарков на каждом используемом языке.
- **Автоматический анализ параллельности кода с помощью эвристик или методов многогранного анализа.** Технологии данного типа предназначены для вычисления зависимостей данных и пространств итераций с помощью точных методов или эвристик. Эвристики в настоящее время являются частью большинства коммерческих компиляторов, когда как в составе открытых и экспериментальных решений можно найти более сложные методы, например, многогранный анализ (polyhedral analysis). В работе [12] для компилятора GCC реализовано расширение для автоматической идентификации параллельных циклов и генерации для них кода на OpenCL. Аналогичное расширение PPCG для компилятора Clang (LLVM) [13] способно преобразовывать код на C/C++ в CUDA-ядра. Обе технологии преобразуют вычислитель-

ные циклы из внутреннего представления компилятора в код на OpenCL или CUDA при помощи системы многогранного анализа Chunky Loop Generator (CLooG) [14]. Source-to-source компилятор Par4all [15] преобразует код на языке C или Fortran в код CUDA, OpenCL или OpenMP с помощью системы многогранного анализа PIPS.

Явное программирование на CUDA, директивные расширения и DSL-языки в любом случае предполагают модификацию или переработку исходного кода программы. По этой причине портирование больших приложений на GPU с помощью этих технологий сильно затруднено. Если же приложение портировано лишь частично, то синхронизация данных между хостом и GPU может значительно влиять на общую производительность. Так, при портировании только одного блока WSM5 модели WRF с помощью директив PGI Accelerator, время обменов данными составляет 40–60 % общего времени [20].

На основе сопоставления свойств существующих технологий с требованиями, возникающими при портировании на GPU типичного вычислительного приложения, можно выделить ряд возможностей, имеющих потенциально наиболее важную роль при планировании и разработке программных систем следующего поколения:

- Поддержка широкого множества *существующих* языков программирования;
- Автоматическая оценка параллельности вычислительных циклов, не требующая внесения изменений в исходный код или каких-либо дополнительных действий со стороны пользователя;
- Генерация кода, полностью совместимая со стандартной хост-компиляцией;
- Минимизация обмена данными между памятью системы и GPU;
- Встраивание в существующие схемы распараллеливания, в первую очередь – MPI.

Целью проекта KernelGen является создание компилятора, удовлетворяющего всем перечисленным условиям и проработка стратегии развития необходимых для этого технологий. Очевидно, что подобная система не может быть построена ни на основе директивных расширений, ни на основе DSL, в то же время в ней вполне могли бы быть использованы наработки исследовательских решений по автоматическому анализу циклов.

Данная статья организована следующим образом. В разделе 1 предлагаются решения по организации процессов компиляции, компоновки и генерации кода, а также нестандартная модель исполнения, позволяющая естественным образом обеспечить более эффективное взаимодействие параллельных частей кода на GPU. В разделе 2 излагается способ модификации существующей технологии анализа параллельности циклов для генерации GPU-кода. Разделы 3 и 4 посвящены, соответственно, необходимым дополнительным подсистемам исполнения приложений и сравнительному анализу работы тестовых задач.

1. Этапы преобразования кода

При разработке системы компиляции на основе существующих наработок значительную роль играет выбор наиболее подходящей базовой инфраструктуры по большому числу критериев: наличие фронтендов для различных языков, полнота и гибкость внутреннего представления, существование базового набора оптимизирующих преобразований и эффективных бекендов для целевых архитектур, динамика развития и поддержка со стороны сообщества разработчиков. Наиболее развиты по этим критериям компиляторы GCC, LLVM и Open64. Компилятор GCC поддерживает наибольшее число языков программирования, но не имеет бекендов для GPU, когда как LLVM и Open64 имеют бекенды для

NVIDIA PTX ISA. Компилятор Open64 имеет фронтенды для C, C++ и Fortran, генерирует качественный код, но при этом, к сожалению, имеет сильно сегментированное сообщество разработчиков, развивающих множество отдельных веток кода в интересах коммерческих компаний и исследовательских организаций. Компилятор LLVM не имеет собственного фронтенда для языка Fortran, но способен при помощи плагина DragonEgg [16] использовать фронтенды компилятора GCC. При этом он имеет собственный GPU-бекенд NVPTX, имеет простое внутреннее представление (LLVM IR – intermediate language) и развивается намного более интенсивно, чем GCC и Open64. Из этих соображений, за основу для KernelGen был выбран LLVM.

Компилятор KernelGen работает напрямую с оригинальным приложением, не требуя каких-либо изменений ни в исходном коде, ни в системе сборки. За счет использования фронтенда незначительно модифицированной версии GCC, он полностью совместим с его опциями, что гарантирует высокий уровень поддержки большого числа приложений. Чтобы обеспечить стандартный процесс сборки, в KernelGen используется схема, напоминающая LTO (link time optimization – инфраструктура компилятора для дополнительной оптимизации кода во время компоновки): код для GPU сначала добавляется в отдельную секцию объектных файлов, затем объединяется и снова разделяется на отдельные ядра на этапе компоновки. Окончательная компиляция GPU-ядер в ассемблер происходит при необходимости, уже во время работы приложения (JIT, just-in-time compilation). Схема основных этапов преобразования кода приведена на рис. 1.

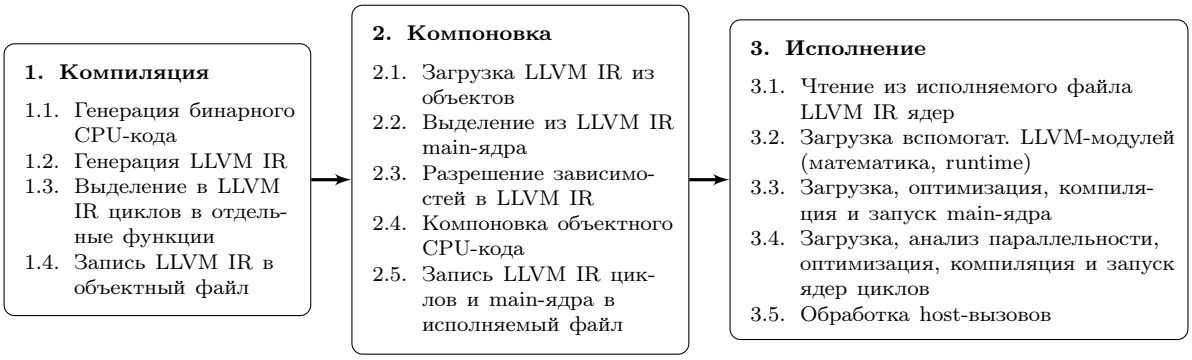

Рис. 1. Этапы преобразования кода компилятором KernelGen

В результате работы компилятора, исходное приложение преобразуется во множество GPU-ядер: одно или несколько *основных* ядер и множество *вычислительных* ядер. Основные ядра исполняются на GPU в одном потоке. Их задача – хранить данные, исполнять небольшие последовательные участки кода и производить вызовы вычислительных ядер и отдельных CPU-функций, которые невозможно или неэффективно переносить на GPU. Вычислительные ядра исполняются на GPU множеством параллельных нитей с полной загрузкой мультипроцессоров. Таким образом, максимальная доля кода выполняется на GPU, а CPU лишь координирует исполнение. В частности, при работе MPI-приложения каждый рабочий процесс в данном случае будет представлять собой GPU-ядро с небольшим числом CPU-вызовов MPI. Использование MPI дополнительно облегчается за счет поддержки GPU-адресов в командах обмена данными [18]. В целом, такая модель исполнения имеет много общего с native-режимом Intel MIC, но работает на GPU, где скалярные вычислительные блоки способны достигать высокой эффективности без необходимости векторизации.

1.1. Компиляция

При компиляции отдельных объектов, генерируется как x86-ассемблер (таким образом, приложение по-прежнему работоспособно при отсутствии GPU), так и представление LLVM IR. Для разбора исходного кода используется компилятор GCC, чье внутренее представление *gimple* преобразуется в LLVM IR с помощью плагина DragonEgg. Затем в IR-коде производится выделение тел циклов в отдельные функции, вызываемые через универсальный интерфейс *kernelgen_launch* (рис. 2), где *kernel* – имя или адрес функции (вместо имен в начале работы программы подставляются адреса), *data* – структура, агрегирующая аргументы вызова, *szdata* и *szdatai* – размер списка аргументов и списка целочисленных аргументов (последний используется для вычисления хеша функции и поиска ранее скомпилированных ядер во время исполнения).

```
__device__ int kernelgen_launch(
    unsigned char* kernel, unsigned long long szdata,
    unsigned long long szdatai, unsigned int* data);
```

Рис. 2. Интерфейс вызова функций-циклов

Стандартный механизм выделения каскадов вложенных циклов в функции LLVM *LoopExtractor* расширен, так чтобы цикл не заменялся, а дополнялся вызовом функции по условию, как показано на рис. 3. С помощью данного условия runtime-библиотека KernelGen может переключать выполнение между различными версиями цикла. Например, если цикл определен как непараллельный, то *kernelgen_launch* возвращает -1, и код цикла начинает выполняться основным ядром в последовательном режиме. Тем не менее, данный цикл может содержать вложенные параллельные циклы, обработка которых будет проведена аналогичным образом. В конце концов, если весь каскад тесно вложенных циклов непараллелен, несовместим (например, содержит вызовы внешних CPU-функций) или оценен как неэффективный для GPU, то вся функция выгружается для работы на хосте с помощью вызова *kernelgen_hostcall* (рис. 4), при котором GPU-приложение останавливает свою работу и передает данные и адрес функции для выполнения на CPU. Функции *kernelgen_launch* и *kernelgen_hostcall* работают в GPU-ядре и вызывают остановку его выполнения. После завершения работы другого ядра или CPU-функции, основное ядро продолжает работу. Хост-часть управляющих функций компилирует и выполняет заданную функцию с помощью интерфейса FFI (Foreign Function Interface).

```
if (kernelgen_launch(kernel, szdata, szdatai, data) == -1) {
    // Запустить оригинальный код цикла.
}
```

Рис. 3. Переключение версий кода цикла между функцией-ядром и оригинальным кодом

```
__device__ void kernelgen_hostcall(
    unsigned char* kernel, unsigned long long szdata,
    unsigned long long szdatai, unsigned int* data);
```

Рис. 4. Интерфейс хост-вызова

Одним из специфических свойств KernelGen является хранение всех данных приложения в памяти GPU. Для того чтобы обеспечить его совместимость с наличием CPU-вызовов, реализована простая система синхронизации памяти. При попытке CPU-функции обратиться к памяти по адресу из диапазона GPU возникающий сигнал сегментации обрабатывается дублированием страниц из памяти GPU в страницы CPU-памяти, расположенные по тем же адресам. После завершения работы CPU-функции, измененные CPU-страницы синхронизируют изменения с памятью GPU.

1.2. Компоновка

Во время компоновки отдельных объектов в приложение или библиотеку, LLVM IR также линкуется в один общий IR-модуль для main-ядра и по одному IR-модулю на каждый вычислительный цикл. IR-код погружается в исполняемый файл и в дальнейшем оптимизируется и компилируется в GPU код по мере необходимости во время работы приложения.

Специальной обработки требуют глобальные переменные. Синхронизация глобальных переменных между ядрами потребовала бы разработки для GPU динамического компоновщика. Вместо этого, в начале работы программы на CPU передаются адреса всех глобальных переменных. Во время выполнения, виртуальные глобальные переменные заменяются на соответствующие фактические адреса. Это корректно, т.к. в LLVM глобальная переменная реализована как указатель на память, содержащую ее логическое значение.

1.3. Модель исполнения

Основное ядро запускается в самом начале выполнения приложения и работает на GPU постоянно. Во время работы вычислительного ядра или CPU-функции основное ядро переходит в состояние активного ожидания и продолжает работу после завершения внешнего вызова. Для реализации данной схемы GPU должно поддерживать одновременное исполнение нескольких ядер (concurrent kernel execution) или временную выгрузку активного ядра (kernel preemption). Одновременное исполнение ядер доступно в GPU NVIDIA, начиная с Compute Capability 2.0, в GPU AMD такой возможности нет, но есть вероятность появления kernel preemption в одной из следующих версий OpenCL. По этой причине в данный момент KernelGen работает только с CUDA.

Вызовы *kernelgen_launch* и *kernelgen_hostcall* состоят из двух частей: device-функции на GPU и одноименного вызова в CPU-коде, который выполняет, соответственно, окончательную генерацию кода и запуск вычислительного ядра или загрузку данных с GPU и запуск CPU-функции средствами Foreign Function Interface (FFI). Взаимодействие между частями может быть организовано посредством глобальной памяти GPU или pinned-памяти хоста. Однако для гарантированной передачи корректного значения необходимо обеспечить *атомарный* режим операций чтения и записи, доступность которого является определяющим фактором. По этой причине был реализован метод, использующий глобальную память.

На GPU Kepler K20 запуск ранее скомпилированых ядер может быть проведен без взимодействия с CPU, непосредственно из основного ядра, с помощью технологии динамического параллелизма.

Дополнительное препятствие взаимодействию GPU-ядра с другим ядром или CPU состоит в том, что данные нити (CUDA thread) хранятся в регистрах или локальной памяти. Это означает, что аргументы, переданные из основного GPU-ядра не могут быть использо-

ваны где-либо, кроме как в нем самом. Для преодоления этого ограничения, бекенд NVPTX изменен так, чтобы локальные переменные помещаются не в *.local*-секцию, а в *.global*, делая их доступными всем GPU-ядрам и хосту.

2. Генерация CUDA-ядер для параллельных циклов

Частью инфраструктуры LLVM является библиотека Polly [2] (от polyhedral analysis – многогранный анализ) – оптимизирующее преобразование циклов, основанное на CLooG. Оно способно распознавать параллельные циклы в IR-коде, оптимизировать кеширование за счет добавления блочности, оптимизировать доступ к памяти за счет перестановки циклов и генерировать код, использующий OpenMP. Для заданного кода CLooG строит *абстрактное синтаксическое дерево* (AST), а затем проводит расщепление циклов по некоторым измерениям. Благодаря возможности расщепления частично-параллельных измерений, для исходного цикла может быть найдено эквивалентное представление из одного или нескольких циклов, часть которых параллельна. Подобный подход используется довольно редко, большинство современных компиляторов ограничиваются проверкой параллельности измерений существующих циклов без глубокого анализа.

Polly работает с частями программы, для которых можно статически (без выполнения) предсказать поток управления и доступ в память в зависимости от фиксированного набора параметров. Такие части принято называть *статическими частями потока управления* (static control parts – SCoPs). Часть программы представляет собой SCoP при выполнении следующих условий:

1. Поток управления формируется условными операторами и циклами-счетчиками:
   (a) каждый цикл-счетчик имеет одну индексную переменную с константным шагом изменения, верхняя и нижняя границы цикла заданы афинными выражениями, зависящими от параметров и индексных переменных внешних циклов;
   (b) условные операторы сравнивают значения двух афинных выражений, зависящих от параметров SCoP и индексных переменных.
2. Обращения в память происходят со смещениями от указателей-параметров SCoP. Смещения задаются афинными выражениями от параметров SCoP и индексных переменных циклов.
3. Содержит вызовы только функций без побочных эффектов.

Первое условие означает структурированность потока управления: код можно логически разбить на иерархию вложенных блоков, имеющих один вход и один выход, каждый блок полностью вложен в объемлющий. Запрещены конструкции, нарушающие структуру потока управления (*break*, *goto*). Использование афинных выражений позволяет применять аппарат целочисленного программирования для расчета границ циклов и обращений в память в зависимости от параметров SCoP.

Если работать с высокоуровневым представлением программы (код на языке высокого уровня, абстрактное синтаксическое дерево), то множество приемов программирования (например, арифметика указателей, циклы *while*, операторы *goto*) будут нарушать описанные требования. Если же перейти к промежуточному, близкому к ассемблеру представлению, каковым является LLVM IR, то арифметика указателей будет реализована как набор арифметических операций с регистрами, и любой цикл, вне зависимости от типа (*for*, *while*),

будет реализован как обычный условный переход. Следствием работы Polly с LLVM IR являются две полезные возможности:

- распараллеливать *while*-циклы, когда как, например, стандартом OpenACC такая возможность не предусмотрена;
- распараллеливать циклы с адресной арифметикой, что, например, не поддерживается в PGI OpenACC.

При адаптации Polly для получения GPU-ядер был использован существующий генератор кода OpenMP, работающий следующим образом. Если внешний цикл является параллельным, то его содержимое перемещается в отдельную функцию, с добавлением вызовов функций библиотеки libgomp – GNU реализации OpenMP. При этом распределение итераций по ядрам производит среда исполнения, а распараллеливается только самый внешний цикл. Для KernelGen в эту логику были внесены следующие изменения:

1. Отображение пространства итераций на нити GPU, с учетом необходимости объединения запросов в память нитей варпа (coalescing transaction);
2. Рекурсивная обработка вложенных циклов с целью использования возможностей GPU по созданию многомерных сеток нитей.

Пусть в заданной группе циклов можно распараллелить N тесно-вложенных циклов. Тогда ядро может быть запущено на решетке с числом измерений N (для CUDA $N \leq 3$). Для каждого измерения, распределяемого между нитями GPU, генерируется код, рассчитывающий положение нити в блоке и блока в сетке. Каждому параллельному циклу ставится во взаимно однозначное соответствие измерение решетки, причем в обратном порядке – внутреннему циклу соответствует измерение X (это позволяет объединять запросы в память). Для каждого параллельного цикла генерируется код, определяющий нижнюю и верхнюю границы части пространства итераций, которая должна быть выполнена нитью. Затем генерируется последовательный код цикла с измененными границами и шагом.

Схема этапов работы Polly, анализа и генерации кода приведена на рис. 5.

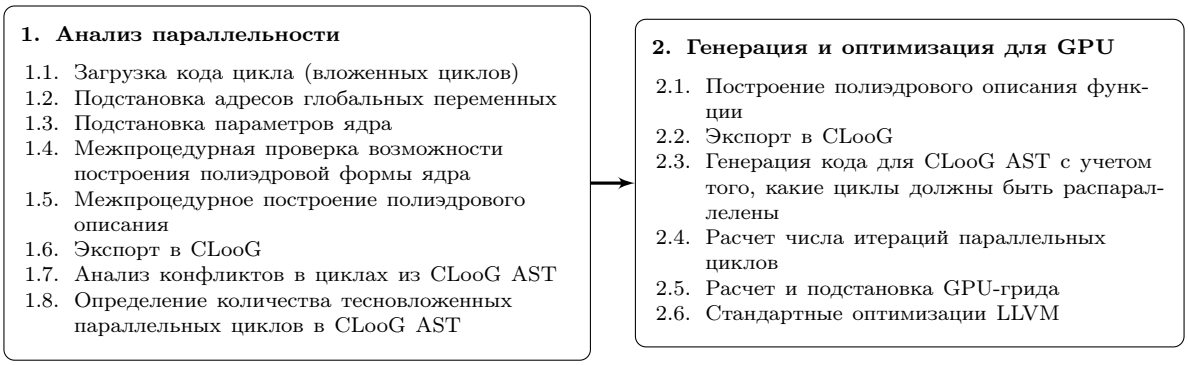

Рис. 5. Этапы генерации CUDA-ядер для параллельных циклов компилятором KernelGen. Оптимизация происходит сразу для всего SCoP, генерация – для каждой функции в отдельности

3. Дополнительные средства времени исполнения

Включение бекенда NVPTX для генерации GPU-кода в LLVM позиционировалось компанией NVIDIA как «открытие компилятора». Тем не менее, помимо того, что закрытым остается C/C++/CUDA-фронтенд, a clang обладает лишь минимальной поддержкой неко-

торых ключевых слов CUDA, недоступной также остается часть компилятора существенно важная для его применения в LLVM: библиотека математических функций C99. Поскольку в рамках закрытого компилятора CUDA эти функции реализованы в виде заголовочных файлов C/C++, их использование с другими языками на уровне LLVM невозможно, и пользователю NVPTX-бекенда доступны только функции, встроенные в аппаратуру (builtins), среди которых, например, нет точных версий функций sin, cos, pow. В KernelGen данная проблема решена путем конвертации заголовочных файлов в LLVM IR одним из двух способов: с помощью clang (требуется множество модификаций, полученный IR-код предположительно содержит некорректные части) или с помощью cicc (вызывается из nvcc). В последнем случае, IR-модуль можно получить, отправив на компиляцию пустой .cu-файл и выгрузив код IR-модуля из cicc с помощью отладчика. IR, сгенерированный cicc совместим с актуальной версией LLVM и позволяет производить компоновку математической библиотеки и использующего ее приложения на уровне IR-кода, вне зависимости от начального языка. В частности, благодаря этому KernelGen позволяет генерировать GPU-код для программ на языке Fortran, использующих любые стандартные математические функции.

Некоторые типы функций CUDA API, такие как выделение GPU-памяти и загрузка другого CUDA-модуля, всегда приводят к неявной синхронизации асинхронных операций. Поскольку схема работы KernelGen требует постоянного поддержания основного ядра в состоянии выполнения, необходимость выделения памяти или загрузки новых ядер в процессе его работы приведет к блокировке. В этом отношении существующие версии CUDA создают для развития KernelGen определенные препятствия, вынуждая реализовывать нестандартные эквиваленты базовой функциональности.

Если синхронность выделения памяти на GPU со стороны хоста еще можно считать разумным ограничением, то подтвержденная экспериментами синхронность вызовов *malloc* внутри GPU-ядер явно избыточна, так как память для индивидуальных потоков выделяется заранее. Так как оба стандартных варианта не могут быть использованы, KernelGen выполняет начальную преаллокацию памяти для собственного динамического пула и управляет его работой.

JIT-компиляция вычислительных ядер предполагает, что вновь скомпилированные GPU-ядра будут загружаться на GPU в фоне работающего основного ядра. Обычно динамическую загрузку ядер можно произвести с помощью стандартных функций *cuModuleLoad* и *cuModuleGetFunction* CUDA Driver API. Однако, обе эти функции являются синхронными, предположительно, из-за неявного выделения памяти для хранения кода и статических данных. В данной ситуации при разработке KernelGen не оставалось иного выбора, кроме как реализовать загрузчик кода новых ядер вручную, предварительно создав для них пустую функцию-контейнер. Загрузчик основан на технологиях проекта AsFermi [3] и действует следующим образом. В начале работы приложение на GPU загружается достаточно больше пустое ядро (содержащее инструкции NOP). По мере того, как в процессе работы приложения требуется запускать вновь скомпилированные ядра, их код копируется как данные в адресное пространство контейнера, которое известно благодаря инструкции LEPC (получить значение Effective Program Counter). Контейнер размещает код множества небольших ядер друг за другом, создавая своеобразный динамический пул памяти для кода. При этом необходмо учитывать, что различные ядра могут использовать различное число регистров. Для этого загрузчик создает 63 фиктивных ядра (точки входа), использу-

ющих от 1 до 63 регистров с единственной инструкцией JMP для перехода по адресу начала требуемого ядра в контейнере.

Система синхронизации памяти между GPU и хостом использует вызов *mmap*, ограничивающий возможные диапазоны адресов величинами, кратными размеру страниц (4096 байт). Поэтому выравнивание всех данных GPU по границе 4096 было бы очень удобным упрощением на данном этапе. К сожалению, текущая реализация CUDA (5.0) учитывает настройки выравнивания данных при компиляции, но при этом игнорирует их во время исполнения. Обход этого дефекта реализован посредством выравнивания размеров всех данных по границе 4096 вручную с помощью функций библиотеки *libelf*.

4. Тестирование

KernelGen тестируется на трех типах приложений: тесты корректности, тесты производительности и работа на реальных задачах. Тесты корректности предназначены для контроля регрессивных изменений в генераторе кода, тесты производительности позволяют анализировать эффективность текущей версии KernelGen в сравнении с предыдущими сборками и другими компиляторами. При тестировании производительности предпочтение отдается сравнению с результатами других распараллеливающих компиляторов, поскольку в отличие от сравнения с кодом, оптимизированным вручную, это позволяет проанализировать достоинства и недостатки компилятора в своем классе систем.

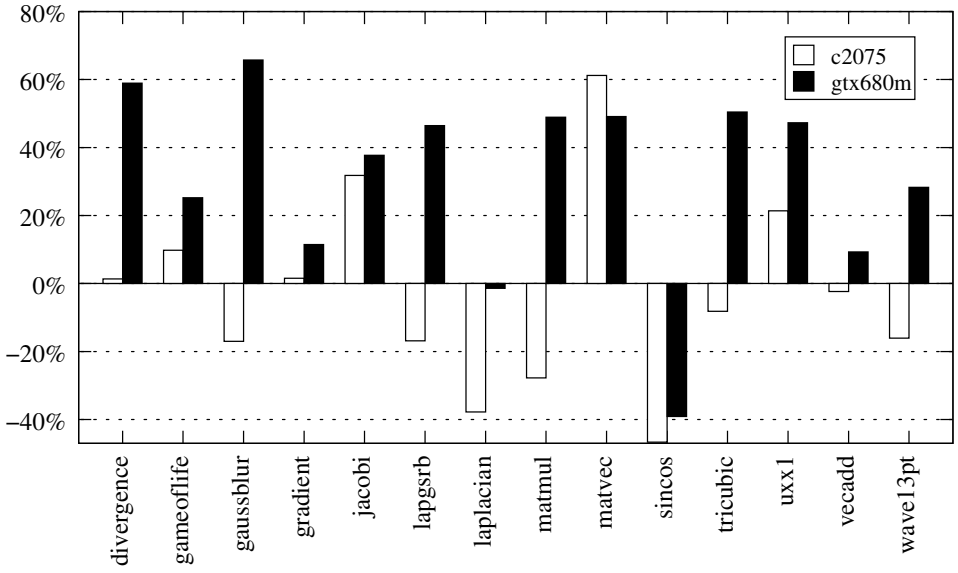

Рис. 6. Сравнение производительности вычислительных ядер некоторых тестовых приложений, скомпилированных KernelGen r1780 и PGI 13.2 на GPU NVIDIA Tesla C2075 (Fermi sm_20) и GTX 680M (Kepler sm_30). Положительные величины – ядра, собранные KernelGen быстрее, чем у PGI, отрицательные величины – ядра, собранные KernelGen медленнее, чем у PGI

Для тестирования были выбраны приложения, реализующие различные типовые алгоритмы на двумерной или трехмерной регулярной сетке с одинарной точностью (часть тестов адаптировано из материалов работы [1], описание и исходный код приведены в [21]). На рис. 6 показано, насколько меняется скорость работы тестов, скомпилированных с помощью KernelGen, по сравнению с версией PGI OpenACC. Соответствующие абсолютные

времена и число регистров для каждой версии теста приведены в таблице. Тесты *jacobi*, *matmul* и *sincos* реализованы на языке Fortran, остальные тесты – на C (также были проверены реализации тестов *wave13pt* и *laplacian* на языке C++, однако для данного сравнения они не пригодны, поскольку компилятор PGI не поддерживает директивы OpenACC в C++). KernelGen автоматически распознает наличие вложенных параллельных циклов внутри непараллельного цикла по числу итераций, когда как PGI делает это только при соответствующей ручной расстановке директив OpenACC. Более низкая производительность KernelGen на тесте *matmul* обусловлена тем, что компилятор PGI реализует частичную раскрутку внутреннего цикла с редукцией на регистрах. Более низкая производительность ряда других тестов требует отдельного изучения.

Таблица

Сравнение времени исполнения (сек) и числа регистров (nregs) для вычислительных ядер некоторых тестовых приложений, скомпилированных KernelGen r1780 и PGI 13.2 на GPU NVIDIA Tesla C2075 (Fermi sm_20) и GTX 680M (Kepler sm_30)

Тест	NVIDIA Tesla C2075				NVIDIA GTX 680M			
	KernelGen		PGI		KernelGen		PGI	
	время	nregs	время	nregs	время	nregs	время	nregs
divergence	0,008462	18	0,008578	40	0,007924	20	0,012595	49
gameoflife	0,010180	21	0,011178	28	0,012555	21	0,015723	32
gaussblur	0,016009	56	0,013286	31	0,014991	51	0,024847	36
gradient	0,010582	21	0,010745	42	0,008855	22	0,009869	51
jacobi	0,007824	24	0,010314	23	0,006324	23	0,008711	30
lapgsrb	0,017307	55	0,014386	61	0,018310	40	0,026813	63
laplacian	0,008150	18	0,005070	39	0,007431	21	0,007325	48
matmul	0,000993	16	0,000717	23	0,000730	16	0,00108	33
matvec	0,016349	16	0,026357	22	0,028514	16	0,042523	25
sincos	0,010796	22	0,005757	26	0,008421	22	0,005129	29
tricubic	0,053462	63	0,049090	63	0,064770	61	0,097440	63
uxx1	0,016479	32	0,020002	59	0,015520	32	0,022861	63
vecadd	0,004837	12	0,004723	24	0,004454	12	0,004869	31
wave13pt	0,011778	34	0,009885	54	0,012461	34	0,015986	60

Тестирование на больших вычислительных приложениях COSMO [19] и WRF [20] показало, что KernelGen способен генерировать корректные исполняемые файлы с поддержкой GPU за разумное время.

Заключение

В проекте KernelGen реализована оригинальная схема автоматического портирования кода на GPU, подходящая для сложных приложений. Не требуя никаких изменений в исходном коде, компилятор переносит на GPU максимально возможную часть кода, включая выделение памяти, тем самым создавая эффективную схему для преимущественно GPU-

вычислений. KernelGen реализует средства автоматического анализа параллелизма циклов, основанные на LLVM, Polly и других проектах, расширяя их поддержкой генерации кода для GPU. Генератор GPU-кода основан на NVPTX-бекенде для LLVM, совместно развиваемом компанией NVIDIA и силами сообщества LLVM. Тестирование показало, что GPU-код, генерируемый KernelGen по своей эффективности сравним с коммерческим компилятором PGI.

Для того чтобы начать использовать компилятор в прикладных задачах, остается реализовать некоторые функциональные элементы. В частности, в нынешней версии отсутствует механизм накопления статистики эффективности исполнения GPU-кода для принятия решений о переключении на CPU-реализацию в неэффективных случаях. В генераторе параллельных циклов желательно добавить возможность использования разделяемой памяти и распознавание в коде идиомы редукции. Запуск вычислительных ядер на архитектуре Kepler может быть организован более эффективно за счет использования динамического параллелизма.

Код KernelGen распространяется по лицензии University of Illinois/NCSA (за исключением плагина для GCC) и доступен на сайте проекта: `http://kernelgen.org/`.

Работа поддержана грантом «Swiss Platform for High-Performance and High-Productivity Computing» (HP2C, `hp2c.ch`), а также контрактами Applied Parallel Computing LLC № 12-2011 и № 13-2011. Тестирование велось на оборудовании, предоставленном компаниями NVIDIA и HP, кластере «Ломоносов» МГУ им М.В. Ломоносова [22] и кластере «Tödi» Швейцарского национального суперкомпьютерного центра (CSCS).

Литература

1. Christen, M. PATUS for Convenient High-Performance Stencils: Evaluation in Earthquake Simulations / M. Christen, O. Schenk, Y. Cui // Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (Salt Lake City, USA, November 10–16, 2012).
2. Grosser, T. Polly – Polyhedral Optimization in LLVM / T. Grosser, H. Zheng, R. Aloor, A. Simbürger, A. Größlinger, L.-N. Pouchet // International Symposium on Code Generation and Optimization (Charmonix, France, April 2–6, 2011).
3. Hou, Y. AsFermi: An assembler for the NVIDIA Fermi Instruction Set / Y. Hou // URL: `http://code.google.com/p/asfermi/` (дата обращения: 03.12.2012).
4. Gysi, T. HP2C Dycore / T. Gysi // Workshop on COSMO dynamical core rewrite and HP2C project — March 2012, Offenbach, Germany — URL: `http://mail.cosmo-model.org/pipermail/pompa/attachments/20120306/079fadc1/DWD_HP2C_Dycore_120305.pdf` (дата обращения: 03.12.2012).
5. Ragan-Kelley, J. Decoupling Algorithms from Schedules for Easy Optimization of Image Processing Pipelines / J. Ragan-Kelley, A. Adams, S. Paris, M. Levoy, S. Amarasinghe, F. Durand. // ACM Trans. Graph. — 2012. — Vol. 31, No. 4. — P. 32:1–32:12.
6. Адинец, А.В. Программирование графических процессов при помощи расширяемых языков. // А.В. Адинец // Вестник Южно-Уральского государственного университета. Серия «Математическое моделирование и программирование». — 2011. — N. 25 (242). — Вып. 9. — С. 52-63.

7. The OpenACC™Application Programming Interface. Version 1.0. URL: http://www.openacc-standard.org (дата обращения: 27.05.2012).
8. OpenHMPP, New HPC Open Standard for Many-Core. URL: http://www.openhmpp.org/en/OpenHMPPConsortium.aspx (дата обращения: 03.12.2012).
9. The Heterogeneous Offload Model for Intel® Many Integrated Core Architecture. URL: http://software.intel.com/sites/default/files/article/326701/heterogeneous-programming-model.pdf (дата обращения: 03.12.2012).
10. Govett, M. Development and Use of a Fortran → CUDA translator to run a NOAA Global Weather Model on a GPU cluster / M. Govett // Path to Petascale: Adapting GEO/CHEM/ASTRO Applications for Accelerators and Accelerator Clusters — 2009. — National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign. — URL: http://gladiator.ncsa.uiuc.edu/PDFs/accelerators/day2/session3/govett.pdf (дата обращения: 27.05.2012).
11. Бахтин, В.А. Автоматическое распараллеливание Фортран-программ на кластер с графическими ускорителями / В.А. Бахтин, Н.А. Катаев, М.С. Клинов, В.А. Крюков, Н.В. Поддерюгина, М.Н. Притула // Параллельные вычислительные технологии (ПаВТ'2012): Труды международной научной конференции (Новосибирск, 26 марта – 30 марта 2012 г.). — Челябинск: Издательский центр ЮУрГУ, 2012. — С. 373–379.
12. Kravets, A. GRAPHITE-OpenCL: Automatic parallelization of some loops in polyhedra representation / A. Kravets, A. Monakov, A. Belevantsev // GCC Developers' Summit (Ottawa, Canada, October 25–27, 2010).
13. Verdoolaege, S. Polyhedral parallel code generation for CUDA / S. Verdoolaege, J. Carlos Juega, A. Cohen, J. Ignacio Gómez, C. Tenllado, F. Catthoor // ACM Trans. Archit. Code Optim. — 2013. — Vol. 9, No. 4. — P. 54:1–54:23.
14. Bastoul, C. Code Generation in the Polyhedral Model Is Easier Than You Think / C. Bastoul // PACT'13 IEEE International Conference on Parallel Architecture and Compilation Techniques (Antibes Juan-les-Pins, France, September 29 - October 3, 2004).
15. Torquati, M. An innovative compilation tool-chain for embedded multi-core architectures / M. Torquati, M. Venneschi, M. Amini, S. Guelton, R. Keryell, V. Lanore, F.-X. Pasquier, M. Barreteau, R. Barrère, C.-T. Petrisor, É. Lenormand, C. Cantini, F. De Stefani. // Embedded World Conference (Nuremberg, Germany, February 2012).
16. Sands, D. Reimplementing llvm-gcc as a gcc plugin / D. Sands // Third Annual LLVM Developers' Meeting. — 2009. — Apple Inc. Campus, Cupertino, California. — URL: http://llvm.org/devmtg/2009-10/Sands_LLVMGCCPlugin.pdf (дата обращения: 03.12.2012).
17. Wolfe, M. The PGI Accelerator Programming Model on NVIDIA GPUs Part 3: Porting WRF / M. Wolfe, C. Toepfer // URL: http://www.pgroup.com/lit/articles/insider/v1n3a1.htm (дата обращения: 03.12.2012).
18. Squyres, J. Open MPI State of the Union / J. Squyres, G. Bosilca, S. Sumimoto, R. vandeVaart // Open MPI Community Meeting. — Supercomputing, 2011. — URL: http://www.open-mpi.org/papers/sc-2011/Open-MPI-SC11-BOF-1up.pdf (дата обращения: 03.12.2012).
19. Consortium for Small-scale Modeling. URL: http://www.cosmo-model.org/ (дата обращения: 03.12.2012).

20. The Weather Research & Forecasting Model. URL: `http://www.wrf-model.org/index.php` (дата обращения: 03.12.2012).

21. KernelGen Performance Test Suite. URL: `https://hpcforge.org/plugins/mediawiki/wiki/kernelgen/index.php/Performance_Test_Suite` (дата обращения: 27.01.2013).

22. Практика суперкомпьютера «Ломоносов» / Вл.В. Воеводин, С.А. Жуматий, С.И. Соболев и др. // Открытые системы. — 2012. — No. 7. — С. 36–39.

Лихогруд Николай Николаевич, аспирант факультета вычислительной математики и кибернетики, МГУ им. М.В. Ломоносова (Москва, Российская Федерация), nicolas@kernelgen.org.

Микушин Дмитрий Николаевич, аспирант, ассистент, Институт информатики Университета Лугано (Швейцария), dmitry@kernelgen.org.

KERNELGEN — A PROTOTYPE OF LLVM-BASED AUTO-PARALLELIZING C/FORTRAN COMPILER FOR NVIDIA GPUs

N.N. Likhogrud, Lomonosov Moscow State University (Moscow, Russian Federation)**,**
D.N. Mikushin, Università della Svizzera italiana (Lugano, Switzerland)

The KernelGen project (`http://kernelgen.org/`) aims to develop Fortran and C compilers based on the state-of-art open-source technologies for automatic GPU kernels generation from unmodified CPU source code, significantly improving the code porting experiences. Parallelism detection is based on LLVM/Polly and CLooG, extended with mapping of loops onto GPU compute grid, and assisted with runtime alias analysis. PTX assembly code is generated with NVPTX backend. Thanks to integration with GCC frontend by means of DragonEgg plugin, and customized linker, KernelGen features full GCC compatibility, and is able to compile complex applications into hybrid binaries containing both CPU and GPU-enabled executables. In addition to more robust parallelism detection, test kernels produced by KernelGen are up to 60 % faster than generated by PGI compiler for kernels source with manually inserted OpenACC directives.

Keywords: GPU, LLVM, OpenACC, JIT-compilation, polyhedral analysis.

References

1. Christen M., Schenk O., Cui Y. PATUS for Convenient High-Performance Stencils: Evaluation in Earthquake Simulations // Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (Salt Lake City, USA, November 10–16, 2012).

2. Grosser T., Zheng H., Aloor R., Simbürger A., Größlinger A., Pouchet L.-N. Polly – Polyhedral Optimization in LLVM // International Symposium on Code Generation and Optimization (Charmonix, France, April 2–6, 2011).

3. Hou Y. AsFermi: An assembler for the NVIDIA Fermi Instruction Set / Y. Hou // URL: `http://code.google.com/p/asfermi/` (accessed: 03.12.2012).

4. Gysi T. HP2C Dycore // Workshop on COSMO dynamical core rewrite and HP2C project. March 2012, Offenbach, Germany. URL: `http://mail.cosmo-model.org/pipermail/pompa/attachments/20120306/079fadc1/DWD_HP2C_Dycore_120305.pdf` (accessed: 03.12.2012).

5. Ragan-Kelley J., Adams A., Paris S., Levoy M., Amarasinghe S., Durand F. Decoupling Algorithms from Schedules for Easy Optimization of Image Processing Pipelines // ACM Trans. Graph. 2012. Vol. 31, No. 4. P. 32:1–32:12.

6. Adinetz A.V. Programming graphics processors with extensible languages // Bulletin of South Ural State University. Series "Mathematical Modelling, Programming and Computer Software". 2011. No. 25(242). P. 52-63.

7. The OpenACC™Application Programming Interface. Version 1.0. URL: `http://www.openacc-standard.org` (accessed: 27.05.2012).

8. OpenHMPP, New HPC Open Standard for Many-Core. URL: `http://www.openhmpp.org/en/OpenHMPPConsortium.aspx` (accessed: 03.12.2012).

9. The Heterogeneous Offload Model for Intel® Many Integrated Core Architecture. URL: `http://software.intel.com/sites/default/files/article/326701/heterogeneous-programming-model.pdf` (accessed: 03.12.2012).

10. Govett, M. Development and Use of a Fortran → CUDA translator to run a NOAA Global Weather Model on a GPU cluster / M. Govett // Path to Petascale: Adapting GEO/CHEM/ASTRO Applications for Accelerators and Accelerator Clusters — 2009. — National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign. — URL: `http://gladiator.ncsa.uiuc.edu/PDFs/accelerators/day2/session3/govett.pdf` (accessed: 27.05.2012).

11. Bakhtin V.A., Kataev N.A, Klinov M.S., Kryukov V.A., Podderyugina N.V., Pritula M.N. Avtomaticheskoe rasparallelivanie Fortran-programm na klaster s graficheskimi uskoritelyami [Automatic parallelization of Fortran programs for GPU-enabled clusters] // Parallel Computational Technologies (PAVT'2012): Proceedings of the International Scientific Conference (Novosibirsk, March 26 – 30, 2012). Chelyabinsk, Publishing of the South Ural State University, 2012. P. 373–379.

12. Kravets A., Monakov A., Belevantsev A. GRAPHITE-OpenCL: Automatic parallelization of some loops in polyhedra representation // GCC Developers' Summit (Ottawa, Canada, October 25–27, 2010).

13. Verdoolaege S., Carlos Juega J., Cohen A., Ignacio Gómez J., Tenllado C., Catthoor F. Polyhedral parallel code generation for CUDA // ACM Trans. Archit. Code Optim. 2013. Vol. 9, No. 4. P. 54:1–54:23.

14. Bastoul C. Code Generation in the Polyhedral Model Is Easier Than You Think / C. Bastoul // PACT'13 IEEE International Conference on Parallel Architecture and Compilation Techniques (Antibes Juan-les-Pins, France, September 29 - October 3, 2004).

15. Torquati M., Venneschi M., Amini M., Guelton S., Keryell R., Lanore V., Pasquier F.-X., Barreteau M., Barrère R., Petrisor C.-T., Lenormand É., Cantini C., De Stefani F. An innovative compilation tool-chain for embedded multi-core architectures // Embedded World Conference (Nuremberg, Germany, February 2012).

16. Sands D. Reimplementing llvm-gcc as a gcc plugin / D. Sands // Third Annual LLVM Developers' Meeting. — 2009. — Apple Inc. Campus, Cupertino, California. — URL: `http://llvm.org/devmtg/2009-10/Sands_LLVMGCCPlugin.pdf` (accessed: 03.12.2012).

17. Wolfe M., Toepfer C. The PGI Accelerator Programming Model on NVIDIA GPUs Part 3: Porting WRF // URL: `http://www.pgroup.com/lit/articles/insider/v1n3a1.htm` (accessed: 03.12.2012).

18. Squyres J., Bosilca G., Sumimoto S., vandeVaart R. Open MPI State of the Union // Open MPI Community Meeting. Supercomputing, 2011. URL: `http://www.open-mpi.org/papers/sc-2011/Open-MPI-SC11-BOF-1up.pdf` (accessed: 03.12.2012).

19. Consortium for Small-scale Modeling. URL: `http://www.cosmo-model.org/` (accessed: 03.12.2012).

20. The Weather Research & Forecasting Model. URL: `http://www.wrf-model.org/index.php` (accessed: 03.12.2012).

21. KernelGen Performance Test Suite. URL: `https://hpcforge.org/plugins/mediawiki/wiki/kernelgen/index.php/Performance_Test_Suite` (accessed: 27.01.2013).

22. Voevodin Vl.V., Zhumatiy S.A., Sobolev S.I. et al. Practice of "Lomonosov" Supercomputer // Open Systems. 2012. No. 7. P. 36–39.

Поступила в редакцию 8 июня 2013 г.